ENVIRONMETRICS
Environmetrics 2002; **13**: 775–798 (DOI: 10.1002/env.557)

Bilinear estimation of pollution source profiles and amounts by using multivariate receptor models

Eun Sug Park[1,*,†], Clifford H. Spiegelman[2] and Ronald C. Henry[3]

[1]*NRCSE, University of Washington, Seattle, WA 98195, and Texas Institute, Texas A&M University, College Station, TX 77843, U.S.A.*
[2]*Texas A&M University, College Station, TX 77843, U.S.A.*
[3]*University of Southern California, Los Angeles, CA 90089, U.S.A.*

SUMMARY

Multivariate receptor models aim to identify the pollution sources based on multivariate air pollution data. This article is concerned with estimation of the source profiles (pollution recipes) and their contributions (amounts of pollution). The estimation procedures are based on constrained nonlinear least squares methods with the constraints given by nonnegativity and identifiability conditions of the model parameters. We investigate several identifiability conditions that are appropriate in the context of receptor models, and also present new sets of identifiability conditions, which are often reasonable in practice when the other traditional identifiability conditions fail. The resulting estimators are consistent under appropriate identifiability conditions, and standard errors for the estimators are also provided. Simulation and application to real air pollution data illustrate the results. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: multivariate receptor model; model identifiability; constrained nonlinear least squares; consistency; bootstrap

1. INTRODUCTION

Multivariate receptor modeling is a collection of methods used to identify pollution sources based on a series of concentrations of airborne gases or particles measured over time. Traditionally, a multivariate receptor model has been applied to the measurements on multiple chemical species (say p different chemical species) collected at a single monitoring site (a receptor). The basic assumption in receptor modeling is chemical mass balance (see, for example, Henry *et al.*, 1984; Hopke, 1985, 1991). That is, the total airborne particulate mass measured at the receptor is a linear sum of the contributions of the individual sources. Let q $(q < p)$ be the number of sources. Based on the chemical mass balance equation aforementioned and the assumption that the relative amounts of the chemical species remain approximately the same as particles/gases travelling from sources to the receptor, a multivariate

*Correspondence to: Eun Sug Park, Texas Transportation Institute, The Texas A&M University System, 3135 TAMU, College Station, TX 77843-3135, U.S.A.
†E-mail: e-park@ttimail.tamu.edu

Contract/grant sponsor: United States Environmental Protection Agency; contract/grant number: CR825173-01-0.
Contract/grant sponsor: National Science Foundation; contract/grant number: DMS-9523878.
Contract/grant sponsor: Center for Environmental Health, Texas A&M University; contract/grant number: P30-ESO9106.

Received 28 November 2000
Accepted 23 May 2001

receptor model takes the form of

$$y_i = \sum_{k=1}^{q} \alpha_{ik} P_k + \varepsilon_i, \qquad i = 1, \ldots, n \tag{1}$$

where $y_i = (y_{i1}, y_{i2}, \ldots, y_{ip})$ is the ith observation at the receptor, $P_k = (p_{k1}, p_{k2}, \ldots, p_{kp})$ is the kth source composition profile consisting of the fractional amount of each chemical species in the emissions from the kth source, α_{ik} is the contribution from the kth source at time i, and $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{ip})$ is the measurement error in the ith observation.

Air pollution data are often obtained as concentrations of a single species (e.g. SO_2) measured from multiple monitoring sites. This can be considered as a multivariate dataset for which the different monitoring sites play the role of different variables. Let p be the number of monitoring sites. Assuming the sources are spatially distinct and environmental conditions (such as wind) over the monitoring sites are fairly stable, (1) can still be applied to this type of data with the spatial profiles substituted for the source composition profiles. Here, the kth source spatial profile represents the relative amounts of the species conveyed to p monitoring sites from the kth source. The closer the monitoring site is to the source, the higher the relative amount in the spatial profile is. Thus, source identification can be facilitated by using these spatial profiles to locate the sources (source regions). One of the purposes of the article is to illustrate how a conventional multivariate receptor model (developed for multivariate air pollution data collected from a single receptor) can be extended to spatial data collected from multiple receptors.

Examples of conventional multivariate receptor modeling include principal component analysis, exploratory factor analysis, target transformation factor analysis, and self-modeling curve resolution (see, for example, Henry, 1991). Estimators from these approaches, however, are not guaranteed to have good statistical properties such as consistency. Also, uncertainty estimates (standard errors) are not provided by those methods.

In matrix terms, model (1) can be written as

$$Y = AP + E \tag{2}$$

where A is an $n \times q$ source contribution matrix, P is an $q \times p$ source (spatial) profile matrix and E is an $n \times p$ error matrix.

In relation to statistical models, this can be viewed as a factor analysis model in the sense that both A and P are the unknown parameters that need to be estimated. (The matrices A and P can be related to a matrix of factor scores and to a factor loading matrix, respectively.) The purpose of factor analysis is to explain the correlations among the observed variables by a set of fundamental factors. It should be noted that our objective in this article is not to predict pollutant concentration at unminitored sites using spatial correlations, but to explain spatial correlations by a set of common factors, i.e. major pollution sources, which is different from the ordinary spatial statistics context. This involves estimating the number of major sources, q, their spatial profiles, P, and contributions, A. This goal, however, cannot be achieved without additional assumptions, which will be given more attention hereafter.

The number of major pollution sources (factors) q will be assumed known throughout the article. Although estimation of q is a nontrivial problem, it is not the purpose of this article (see Henry *et al.*, 1999 and references therein for various practical methods of estimating q).

Even with the known q, there is a fundamental indeterminacy in model (2), i.e. the parameterization is not unique (the model with A and P is equivalent to the model with $A^* = AR$ and $P^* = R^{-1}P$, where

R is a nonsingular $q \times q$ matrix), and so the unique solution to the estimating equation is not guaranteed. To get around this problem, additional assumptions on the parameters (called 'identifiability conditions') are often made in factor analysis such as assuming that $P\Sigma^{-1}P'$ is diagonal (where Σ is the error covariance matrix) or the first q columns of P form an upper triangular matrix under an orthogonal factor model, requiring the lower square submatrix of P to be the identity, or preassigning zeros in specified positions of P (see, for example, Anderson, 1984). Although there could be infinitely many different identifiability conditions in principle, not all of them are physically meaningful or make sense in a given context. Among the conditions aforementioned, preassigning zeros in specified positions of P are often used in receptor modeling. Zero elements in a specified position of P imply that some particular variables do not depend on some specific factors, i.e. some sites (or species if multiple species are measured at a single receptor) are not contributed to by a particular source. This is often a realistic assumption, but not always. In some cases, we may not have any prior information on the source profiles P to determine which source does not contribute to particular sites (or species), or simply there may not be enough zeros in the source profiles to satisfy identifiability conditions.

The following example on air pollution spatial data illustrates this case. The ambient measurements on PM2.5 (the airborne particulate matter less than 2.5 μm in aerodynamic diameter) were collected from 11 monitoring sites in the nearby Grand Canyon National Park during the summer of 1992. The resulting data set consists of 53 observations on 11 variables (receptor sites). A major constituent of PM2.5 is often sulfate formed in the air by oxidation of sulfur dioxide gas. Physically, there are three major source regions of sulfur dioxide gas (which is oxidized to particulate sulfate) in the region: urban and agricultural sources in southern California, a few large electric power plants, and smelters in southern Arizona and northern Mexico. During the summer, the wind patterns are such that only one large power plant is likely to affect the study region. It is located near the center of the region on the map in Figure 1 where CA, AZ and NV all come together. At times, southerly winds during the summer bring moisture and pollution up from the southeast into the study area. These so-called monsoon winds carry airborne sulfate pollution from smelters in southeast AZ and northern Mexico. Source profiles for them were not available, and so traditional identifiability conditions of preassigning

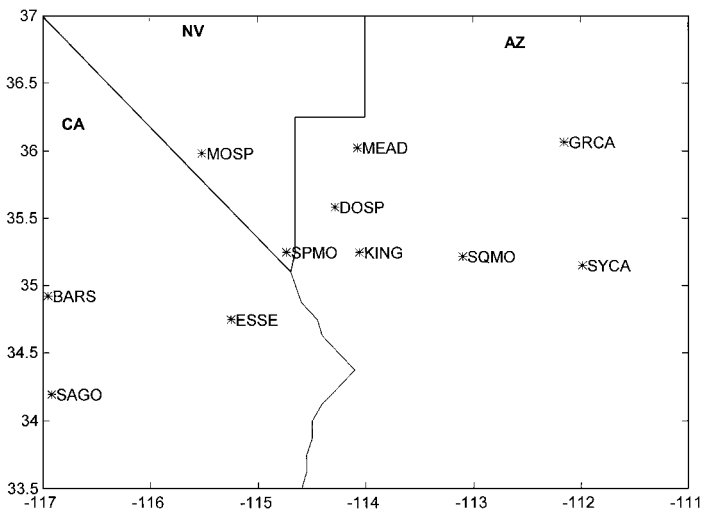

Figure 1. Map of receptor sites (*) near Grand Canyon

zeros in specified positions of P matrix cannot be utilized here. In this case we lack alternative identifiability conditions in the literature. We will revisit this example in Section 4.2. One of the goals of this article is to provide alternative sets of identifiability conditions when the conventional identifiability conditions fail to be satisfied. This issue will be more extensively discussed in Section 2.

The number of parameters in model (2) increases to infinity as the sample size increases. Kiefer and Wolfowitz (1956) showed that without making additional assumptions consistent estimation of the structural parameter (P in our model) is not possible when there are infinitely many incidental parameters (α_i, the rows of A). To get around this problem, they assumed that the incidental parameters were independently distributed chance variables with a common unknown distribution function. This type of model is referred to as a 'structural model (random factor model)' in the literature as opposed to a 'functional model (nonrandom factor model)' that considers α_i to be a vector of nonrandom quantities that varies from one observation to another. The assumption of the structural model was made in some previous works by statisticians in receptor modeling (Bandeen-Roche and Ruppert, 1991; Yang, 1994).

As a matter of fact, many environmental engineers want to view the source contributions as fixed parameters, not random variables, which calls for the functional model. To achieve a consistent sequence of estimators, a parameter space for A, however, needs to be further restricted for the reasons in Kiefer and Wolfowitz (1956). Variations of functional models that restrict the space of incidental parameters, including Quasi Random Functional Model (which is a generalization of the model used in Kiefer and Wolfowitz, 1956) and the Replicated Functional Model, were reviewed and summarized in Gleser (1983).

Quasi-random functional models assume that the first and the second sample moments of the rows of A converge to some fixed vector and matrix, respectively, while treating the rows of A as fixed parameters. With appropriate identifiability conditions, e.g. prespecification of zeros in P, we can get consistent estimates of source profile matrix P and mean contribution $\alpha_0 = \lim_{n\to\infty}\bar{\alpha} = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\alpha_i$, where $\alpha_i = (\alpha_{i1}, \ldots, \alpha_{iq})$ by, for example, constrained nonlinear least squares (Park *et al.*, 1999). Asymptotic normality of the estimator of the source profile matrix can also be established (see Park *et al.*, 1999). The estimates of the individual source contributions, α_i, however, are not consistent since there is only one observation for each contribution.

Replicated functional models assume that there are replicated observations for each α_i. This assumption often makes sense in the context of air pollution study. Air pollution measurements are usually measured over time. Modern instruments are capable of measuring pollution every few fractions of a second. Typically these replicate measurements are aggregated into larger time blocks, for example hourly or daily measurements. The belief and hope of the measurement scientists is that the pollution is stable enough over the aggregation period that aggregation decreases measurement error but does not mask trends. Another example is that the amount of pollution may show some daily (or weekly or monthly) pattern. In that case, it would be more appealing to make an assumption that the source contributions are repeated daily (or weekly or monthly) rather than to treat them all differently. Further, replicated functional models are sometimes more advantageous than quasi-random functional models or structural models because they can, at least, provide consistent estimators not only for the mean contribution but also, for example, for 24 h contributions. In this article, we focus our attention on replicated functional models.

Currently the EPA is considering recommending a standard approach for receptor modeling; see http://www.epa.gov/oar/oaqps/pams/analysis/receptor/rectxtsac.html. The goal of this article is to illuminate the multivariate receptor modeling from a statistical science viewpoint, with the intention of providing methods for estimating the source profiles and contributions consistently, and also to present

new sets of identifiability conditions that are often reasonable in practice. The rest of the article is organized as follows. In Section 2, the issue of model identifiability is revisited, and new sets of identifiability conditions that are often realistic in the context of receptor modeling are presented. Section 3 describes the estimation methods. In Section 4, we illustrate our methods using both simulated data and real data. Finally, closing comments are made in Section 5.

2. IDENTIFIABILITY OF THE MODEL PARAMETERS

We assume that in model (2) each row of matrix E has mean vector 0 and variance–covariance matrix Σ, and A and P are unknown constant matrices. We also place physical constraints on A and P. The elements of A and the elements of P are nonnegative.

$$\alpha_{ik} \geq 0, \quad p_{kj} \geq 0 \tag{3}$$

where $i = 1, \ldots, n, \quad k = 1, \ldots, q, \quad j = 1, \ldots, p$.

We first need to introduce the definition of the model identification.

Definition 1: Let Y be a matrix of the observable random variables, θ be a matrix of the parameters of interest, and $F_Y(C;\theta)$ be the distribution function of Y for parameter θ evaluated at $Y = C$. The parameter θ is identified if, for any θ_1 and θ_2 in the parameter space,

$$F_Y(C;\theta_1) = F_Y(C;\theta_2) \text{ for all } C$$

implies that

$$\theta_1 = \theta_2$$

If the parameter θ is identified, we also say that the model is identified.

A normal distribution plays the least favorable role in model identifiability in the sense that if the parameters are identified under the normal distribution then they are typically identifiable under other distributions but the reverse does not hold (Moran, 1971; Gleser, 1983, 1991). Thus, if we could show identifiability of the parameters under normal distribution (which is entirely determined by its mean and variance), identifiability will in general hold under any other distributions (having mean and variance). In normal error case, the distribution of Y is entirely determined by AP and Σ. That is, $F_Y(C;A_1P_1,\Sigma_1) = F_Y(C;A_2P_2,\Sigma_2)$ implies that $A_1P_1 = A_2P_2$ and $\Sigma_1 = \Sigma_2$, and vice versa. It does not, however, automatically imply that $(P_1, A_1) = (P_2, A_2)$, which are the parameters of our interest. Thus, in our case, Definition 1 can be reduced to the following.

Definition 2: The parameter (P,A) is identified if, for any (P_1, A_1) and (P_2, A_2) in the parameter space,

$$A_1P_1 = A_2P_2$$

implies that

$$P_1 = P_2 \quad \text{and} \quad A_1 = A_2$$

We also define near identifiability of the model parameters.

Definition 3: The parameter (P,A) is *nearly* identified if, for any (P_1, A_1) and (P_2, A_2) in the parameter space,

$$A_1P_1 = A_2P_2$$

implies that

$$P_1 \approx P_2 \quad \text{and} \quad A_1 \approx A_2, \text{elementwise.}$$

That is, for example, the elements of P and A are unique up to three significant digits even if an infinite number of solutions are possible for all the remaining digits.

Since both A and P are unknown, the parameterization for the mean is not unique, i.e. $\mathrm{E}(Y) = AP = A^*P^*$, and so our model is not identified. Under the additional assumption that $\mathrm{rank}(A) = q$, $\mathrm{rank}(P) = q$, it can be shown that $AP = A^*P^*$ always implies that $A^* = AR$ and $P^* = R^{-1}P$ for a nonsingular matrix R. Thus, with the additional rank assumption, nonidentifiability of model (2) is again reduced to the so-called 'factor indeterminacy' problem in factor analysis. Since there are q^2 elements in the matrix R, we need to put q^2 independent conditions on P or A to rule out this indeterminacy. One set of such conditions is:

C1. There are at least $q - 1$ zero elements in each row of P.
C2. The rank of $P^{(k)}$ is $q - 1$, where $P^{(k)}$ is the matrix composed of the columns containing the assigned 0s in the kth row with those assigned 0s deleted.

These conditions can be found in usual multivariate analysis textbooks (see, for example, Anderson, 1984, Section 14.2.2). Under C1–C2, the rows of P are identified except for multiplication by a scale constant (i.e. R is diagonal). By adding a normalization constraint such as

C3-1. $p_{kj} = 1$ for some $j(j = 1, \ldots, P)$ for each $k = 1, \ldots, q$
or
C3-2. $\sum_{j=1}^{p} p_{kj} = 1$ for each $k = 1, \ldots, q$

the multiplication of a row of P by a scale constant can easily be eliminated (i.e. $R = \mathbf{I}$). Factor analysis with these identification conditions (C1, C2, C3-1) has been referred to as 'confirmatory' factor analysis in the literature. Note that the normalization constraint (C3-1 or C3-2) is somewhat arbitrary and does not recover the absolute values in P (in fact, without extra information such as the total mass of each source profile, it is not possible to know the absolute values in P). The constraint $\sum_{j=1}^{p} p_{kj} = 1$ indicates that only the relative amount of species at each site emitted from a source is of our interest. As long as the relative amounts of species in a source profile are given, we consider the source identified. The related but much stronger set of conditions is:

D1. There are at least q columns in P with each of the q columns containing only one nonzero element.
D2. Same as C2.
D3. Same as either C3-1 or C3-2.

These correspond to the assumption of having a tracer element for each source, which have been used for a long time for identifiability in the receptor modeling community. A 'tracer element' is a single species (a single site in this article) that is contributed by a single pollution source, i.e. the nonzero element in D1. It can be easily checked that C1 is automatically satisfied if we have a tracer element for each source, but not vice versa.

Both sets of identifiability conditions (C1–C3 and D1–D3) are the assumptions on the source profile matrix P, and require some prior knowledge about P to get an idea of where to assign 0s. Those conditions were used in some of the previous works in receptor modeling, e.g. D1–D3 by Bandeen-Roach and Ruppert (1991) and Spiegelman and Dattner (1993), and C1–C3 by Yang (1994). Although the conditions C1 and C2 are often reasonable assumptions in receptor modeling, there are some cases

that those conditions cannot be used due to the reasons mentioned in Section 1. Here, we present two new sets of assumptions for identifiability to help solve a factor indeterminacy problem. We need only one set of assumptions to hold for A and P to be identifiable. The first set of our assumptions is:

A1. There are at least $q-1$ zero elements in each column of A.
A2. The rank of $A^{(k)}$ is $q-1$, where $A^{(k)}$ is the matrix composed of the rows containing the assigned 0s in the kth column with those assigned 0s deleted.
A3. $\sum_{j=1}^{p} p_{kj} = 1$ for each $k = 1, \ldots, q$.

The conditions A1–A3 are parallel to the conditions C1–C3, and model identifiability by these conditions can be proved in the same manner (the proof is given in the Appendix). The condition A1 is also closely related to Henry's assumption that the data contain some points such that each source is missing (Henry, 1997a). He argued that if there are at least $(q-1)$ edge points (data points that have one source missing) for each source and the edge points do not have any multicollinearities of dimension less than $q-1$, then the solution to the general mixture problem is unique. Although his conditions are given in terms of data, they can be converted to identifiability conditions on parameters in no error case (in that case, the requirement of $q-1$ edge points is equivalent to A1).

The second set of our assumptions is:

B1. Each source is missing on some days (and we know when a source is missing).
B2. The average contribution of jth $(j = 1, \ldots, q)$ source when the kth $(k \neq j)$ source is missing is equal to the average contribution of the jth source for all days.
B3. The row sum of P is 1.

Define I_k to be a subset of $\{1, 2, \ldots, n\}$ for which the kth source is missing, $\bar{\alpha}_j^{(k)}$ as the average contribution of the jth source when the kth source is missing $(j \neq k)$, and $\bar{\alpha}_j$ as the average contribution of the jth source for all days. Then the above assumptions can be expressed as follows:

For B1–B2,

B1. $\alpha_{ik} = 0$ when $i \in I_k, \qquad k = 1, \ldots, q.$
B2. $\bar{\alpha}_j^{(k)} = \bar{\alpha}_j, \qquad j = 1, \ldots, q, \quad j \neq k.$
B3. $\sum_{j=1}^{p} p_{kj} = 1$ for each $\qquad k = 1, \ldots, q.$

Remark 1: Note the assumption that $\text{rank}(A) = q$ and $\text{rank}(P) = q$ is necessary to make the problem of 'nonunique parameterization of mean' equivalent to the 'factor indeterminacy' problem. Though it is suppressed, the rank assumption on A and P is necessary for all of the identifiability conditions in this section to serve as model identifiability conditions not just as conditions to remove factor indeterminacy.

Remark 2: The identifiability conditions A1–A2 or B1–B2 are physically meaningful. For instance, a factory may shut down for a few days due to an equipment failure and/or repair. The other pollution sources, however, would emit what they have been emitting no matter whether the factory shuts down or not. Also, local weather conditions, say thunderstorms, around the pollution sources in Utah, southern California or Arizona keep the local pollution from reaching the receptors but have no effect on pollution in other areas. That is, the absence of a source at the receptors is due to weather conditions at the source and not at the receptors or at the other sources.

Remark 3: In general, A1–A2 will be handy when the number of sources q is small, whereas B1–B2 will be when q is moderately large—say $q = 7$. For A1–A2 to be satisfied, we need to have at least six zeros for each of seven columns of the A matrix (and the submatrix consisting of the rows containing

zeros with those zeros deleted should be of rank 6). Suppose that one of the sources is missing only for 4 days, which violates A1. In this case, B1 is still satisfied and B2 might also be satisfied.

The following results show that under each set of assumptions, A1–A3 or B1–B3, nonidentifiability of the model parameters can be removed. That is, $A^* = A$ and $P^* = P$. The proofs are given in Appendix A.

Result 1: Let Assumptions A1–A3 hold. Then

$$R = \mathbf{I}$$

where $\mathbf{I}$ is the $q \times q$ identity matrix and R is any nonsingular matrix satisfying $A^* = AR$ and $P^* = R^{-1}P$.

Result 2: Let Assumptions B1–B3 hold. Then

$$R = \mathbf{I}$$

where $\mathbf{I}$ is the $q \times q$ identity matrix and R is any nonsingular matrix satisfying $A^* = AR$ and $P^* = R^{-1}P$.

Remark 4: Condition B2 can be relaxed to $\bar{\alpha}_j^{(k)} \approx \bar{\alpha}_j$, and in that case the parameters are nearly identified (Definition 3), i.e. $R \approx \mathbf{I}$, elementwise.

Remark 5: We emphasize that all the conditions cited in this article are sufficient conditions but not necessary conditions for model identifiability.

3. ESTIMATION OF SOURCE PROFILES AND CONTRIBUTIONS

In Section 3.1, the basic physical model (1) is rewritten in terms of a replicated functional model, and in Section 3.2 an estimation method based on this model is discussed.

3.1. Model

Assume that there are m_i replications for each source contribution α_i. Consider the model

$$\underline{y}_{ij} = \alpha_i P + \underline{\varepsilon}_{ij}, \qquad i = 1, \ldots, n, \quad j = 1, \ldots, m_i \tag{4}$$

where $\underline{y}_{ij}$ (p-dimensional row vector) is the jth replication of the measurement in time period i (for example, $\underline{y}_{11}$ is the first replicate of the measurement on day 1, $\underline{y}_{12}$ is the second replicate of the measurement on day 1, etc.), α_i (q-dimensional row vector) is the source contribution in time period i (the ith source contribution) and $\underline{\varepsilon}_{ij}$ (p-dimensional row vector) is a random error associated with the jth replication of the measurement in time period i. We assume that the $\underline{\varepsilon}_{ij}$s are independent and identically distributed with mean vector $\mathbf{0}$ and covariance matrix $\Sigma = \sigma^2 \mathbf{I}_p$. The source contributions α_i and the source profile matrix P are unknown parameters.

Let $N = \sum_{i=1}^{n} m_i$. In matrix terms, model (4) can be expressed as

$$Y = UAP + E \tag{5}$$

where

$$A = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}, \; U = \begin{bmatrix} \underline{1}_{m_1} & \underline{0} & \cdots & \underline{0} \\ \underline{0} & \underline{1}_{m_2} & \cdots & \underline{0} \\ \vdots & \vdots & \ddots & \vdots \\ \underline{0} & \underline{0} & \cdots & \underline{1}_{m_n} \end{bmatrix}$$

$\underline{1}_{m_i}$ is an m_i-dimensional column vector consisting of 1s, and E is an $N \times p$ error matrix consisting of $\underline{\varepsilon}_{ij}$. We have

$$E(Y) = UAP$$

and

$$\mathrm{Var}(Y) = \mathbf{I}_N \otimes \Sigma$$

Note that U is a known $N \times n$ matrix. Under the identifiability conditions A1–A3 or B1–B3 described in Section 2, $UA_1P_1 = UA_2P_2$ implies that $A_1 = A_2$ and $P_1 = P_2$.

3.2. Estimation

Since we are not making any distributional assumptions, we consider the least squares approach minimizing the sum of squares between the observed value of Y and its mean subject to the constraints on the parameters (nonnegativity and identifiability conditions). This estimation procedure will be referred to as 'constrained nonlinear least squares (CNLS)' hereafter.

The CNLS estimators of A and P are obtained by minimizing the sum of squares,

$$\begin{aligned} Q_N(P,A) &= N^{-1}\mathrm{tr}\big[(Y-UAP)^t(Y-UAP)\big] \\ &= N^{-1}\mathrm{tr}\big[(Y-U\bar{Y})^t(Y-U\bar{Y})\big] + N^{-1}\mathrm{tr}\big[(U\bar{Y}-UAP)^t(U\bar{Y}-UAP)\big] \\ &= N^{-1}\mathrm{tr}\big[(Y-U\bar{Y})^t(Y-U\bar{Y})\big] + N^{-1}\mathrm{tr}\big[M(\bar{Y}-AP)(\bar{Y}-AP)^t\big] \end{aligned} \tag{6}$$

where

$$M = U'U = \begin{bmatrix} m_1 & & & 0 \\ & m_2 & & \\ & & \ddots & \\ 0 & & & m_n \end{bmatrix} \quad \text{and} \quad \bar{Y} = \begin{bmatrix} m_1^{-1}\sum_{j=1}^{m_1} \underline{y}_{1j} \\ m_2^{-1}\sum_{j=1}^{m_2} \underline{y}_{2j} \\ \vdots \\ m_n^{-1}\sum_{j=1}^{m_n} \underline{y}_{nj} \end{bmatrix}$$

over the feasible set Θ,
where

$$\Theta = \{(P,A) | \alpha_{ik} \geq 0, \;\; p_{kj} \geq 0, \;\; i = 1,\ldots,n, \;\; k = 1,\ldots,q, \;\; j = 1,\ldots,p, \;\; r(A) = r(P) = q, \;\; \mathrm{I_A}\}$$

where $r(A) = \mathrm{rank}(A)$, $r(P) = \mathrm{rank}(P)$ and $\mathrm{I_A}$ (=A1–A3 or B1–B3) is a set of the identifiability conditions defined in Section 2. Since the first term of (6) does not depend on A or P, minimizing

$Q_N(P,A)$ is equivalent to minimizing

$$Q_N^*(P,A) = N^{-1}\mathrm{tr}\left[M(\bar{Y}-AP)(\bar{Y}-AP)^t\right]$$

w.r.t. A and P. Since both A and P are unknown parameters, one might think of alternating regression between A and P. Alternating regression, however, may take a huge amount of time to converge or it may not converge at all in some cases. Thus we use the least squares formula for either A or P as a function of the other to get convergence and also to save the running time. According to which set of identifiability conditions is available, we may choose between two types of the fitting algorithms.

Remark 6: When ε has a covariance matrix Σ that is not equal to $\sigma^2\mathbf{I}_p$, we consider the transformed model

$$Y^* = UAP^* + E^*$$

where $Y^* = Y\Sigma^{-\frac{1}{2}}$, $\quad P^* = P\Sigma^{-\frac{1}{2}}$ and $E^* = E\Sigma^{-\frac{1}{2}}$ with Σ known. In practice, laboratories often report the measurement error variance estimates along with the data based upon calibrations or laboratory experience with the instrument. Note that these estimates are obtained from outside the data modeled by (4), and can be used in place of Σ to transform the data matrix. Then (6) can still serve as a valid objective function with Y and P replaced by Y^* and P^*, respectively.

Fitting procedure (CNLS1)

1. Given A, P can be estimated by

$$\tilde{P} = \left[(UA)^t(UA)\right]^{-1}(UA)^tU\bar{Y} = (A^tMA)^{-1}A^tM\bar{Y}$$

2. Find $\tilde{A}$ which minimizes

$$\begin{aligned} Q_N^*(P,A) &= N^{-1}\mathrm{tr}\left[M(\bar{Y}-A\tilde{P})(\bar{Y}-A\tilde{P})^t\right] \\ &= N^{-1}\mathrm{tr}\left[M\left(\bar{Y}-A(A^tMA)^{-1}A^tM\bar{Y}\right)\left(\bar{Y}-A(A^tMA)^{-1}A^tM\bar{Y}\right)^t\right] \\ &= N^{-1}\mathrm{tr}\left[\bar{Y}^tM\left(I_n-A(A^tMA)^{-1}A^tM\right)\bar{Y}\right] \end{aligned}$$

over the feasible set Ω_A, where

$$\Omega_A = \left\{A|\alpha_{ik}\geq 0, \quad i=1,\ldots,n, \quad k=1,\ldots,q, \quad \mathrm{rank}(A)=q, \mathrm{I}_A^*\right\}$$

where I_A^* is a set of the identifiability conditions, A1–A2 or B1–B2, defined in Section 2.

3. Set $\hat{P}$ as $\hat{P} = D^{-1}(\tilde{A}^tM\tilde{A})^{-1}\tilde{A}^tM\bar{Y}$ and $\hat{A}$ as $\hat{A} = \tilde{A}D$, where D is the diagonal normalizing constant matrix with each diagonal element being the row sum of $(\tilde{A}^tM\tilde{A})^{-1}\tilde{A}^tM\bar{Y}$.

Note that Step 3 is needed only to incorporate the normalization constraint. The objective function is as before because DD^{-1} cancels. Since AD is in the space of feasible As, the optimum at each step is exactly the same as it was.

Since we have replicated observations for each α, the resulting estimator for the source contribution matrix, $\hat{A}$, can be shown to be consistent when $m_i \to \infty$, $\lim\frac{m_i}{N} = c_i > 0$.

Theorem (*Consistency of* $\hat{A}$): Let A_0, P_0 and Σ_0 be the true values of A, P and Σ, respectively. Let the parameter space for A, Ω_A be a compact subset of $n \times q$-dimensional Euclidean space containing A_0. Assume the identifiability conditions A1–A3 or B1–B3 in Section 2 are satisfied. Also assume that A_0P_0 is in the interior of a subset of $n \times p$-dimensional Euclidean space. Then, when $m_i \to \infty$, $\lim \frac{m_i}{N} = c_i > 0$, $i = 1, \ldots, n$,

$$\hat{A} \xrightarrow{p} A_0$$

Proof: The proof is given in Appendix B.

It can also be shown that, for the free parameters of A (i.e. after the restrictions on A due to identifiability conditions are taken into account), the corresponding elements of $\hat{A}$ are asymptotically normal assuming that the true parameter values are not on the boundary of the parameter space, i.e. the true parameter values of the free elements of A are nonzero (Park *et al.*, 1999). If the true values of some of the free elements of A are zero, then the limiting distribution of $\hat{A}$ would not be a normal distribution. It would be a mixture of point mass at zero and a normal distribution.

Let $\hat{P} = (\hat{A}^t M \hat{A})^{-1} \hat{A}^t M \bar{Y}$. Consistency of $\hat{P}$ follows from consistency of $\hat{A}$ and the fact that $\bar{Y} \xrightarrow{p} A_0P_0$. Since $\hat{P}$ is a nonlinear function of $\hat{A}$ (and $\bar{Y}$), the asymptotic covariance of $\hat{P}$ is not given in a simple form, though $\hat{P}$ can be proved asymptotically normal. It is well known that in nonlinear least squares the asymptotic variance determined by the delta method is optimistic. This is noted in a number of places, for example, Efron (1982) and Bates and Watts (1988). Also, the errors are more likely to be heavy tailed and cause the normal approximation to require unrealistic sample sizes (here, the number of replications) even if the limiting distribution were normal. For these reasons, we employ a bootstrap method to obtain the approximate covariance matrix of $\hat{P}$. The estimates are asymptotically unbiased by the consistency of $\hat{P}$. We adapt an algorithm in Davison and Hinkley (1997) for our model (5):

For $b = 1, \ldots, B$ (B is the bootstrap size):

1. Find $\hat{A}$ and $\hat{P}$ based on the original data.
2. Compute residuals by $\mathbf{r} = (I - D_H)^{-\frac{1}{2}}(Y - U\hat{A}\hat{P})$, where D_H is a diagonal matrix consisting of the diagonal elements of $H = (U\hat{A})\{(U\hat{A})^t(U\hat{A})\}^{-1}(U\hat{A})^t$.
3. Randomly sample ε_j^* from $r_1 - \bar{r}, \ldots, r_N - \bar{r}$, where r_i $(i = 1, \ldots, N)$ is the ith row of the matrix $\mathbf{r}$ and $\bar{r}$ is the residual mean vector, i.e. $\bar{r} = N^{-1}(\sum_{i=1}^N r_{i1}, \ldots, \sum_{i=1}^N r_{ip})$.
4. Set $Y^* = U\hat{A}\hat{P} + E^*$, where $E^* = (\varepsilon_1^*, \ldots, \varepsilon_N^*)^t$.

Bootstrap estimators $\hat{P}^*$ are obtained for B bootstrap samples, and bootstrap standard errors (BSEs) of $\hat{P}$ can be obtained using the sample covariance matrix of those $\hat{P}^*$.

Remark 7: Though the above fitting procedure is general in the sense that either A1–A3 or B1–B3 can be incorporated in it, especially in the case of A1–A3 being used, we may employ an alternative procedure to reduce a computational burden. The alternative algorithm uses a least squares formula for A as a function of P, and then optimization is done over P instead of over A:

CNLS based on A1–A3 (CNLS2) Let $I_0 = \{1, \ldots, n\} - \cup_{k=1,\ldots q} I_k$, where I_k is defined in Section 2 (i.e. index set when the kth source is missing). Also define A_{I_0} to be the source contribution matrix consisting of the rows corresponding to I_0 and $\alpha_{i(k)}$ to be the ith source contribution vector with the kth element deleted.

1. Given P, A can be estimated by

$$\tilde{A}_{I_0} = \bar{Y}_{I_0} P^t (PP^t)^{-1}$$
$$\tilde{\alpha}_{(k)i} = \bar{y}_i P_{(k)}^t \left(P_{(k)} P_{(k)}^t\right)^{-1}, \qquad i \in I_k$$

where $\bar{Y}_{I_0}$ is the submatrix of $\bar{Y}$ composed of the rows corresponding to I_0 and $P_{(k)}$ is a source composition matrix with the kth row deleted.

2. Find $\tilde{P}$ which minimizes

$$Q_N^*(P, \tilde{A}) = N^{-1} \mathrm{tr}\left[M_{I_0}\left(\bar{Y}_{I_0} - \tilde{A}_{I_0} P\right)\left(\bar{Y}_{I_0} - \tilde{A}_{I_0} P\right)^t\right] + \sum_{i \in I_k,\, k=1,\ldots,q} \frac{m_i}{N}\left(\bar{y}_i - \tilde{\alpha}_{(k)i} P_{(k)}\right)^t \left(\bar{y}_i - \tilde{\alpha}_{(k)i} P_{(k)}\right)$$
$$= N^{-1} \mathrm{tr}\left[M_{I_0} \bar{Y}_{I_0}\left(I_p - P^t (PP^t)^{-1} P\right) \bar{Y}_{I_0}^t\right] + \sum_{i \in I_k,\, k=1,\ldots,q} \frac{m_i}{N} \bar{y}_i \left(I_p - P_{(k)}^t \left(P_{(k)} P_{(k)}^t\right)^{-1} P_{(k)}\right) \bar{y}_i^t$$

over the feasible set Ω_p for which the nonnegativity constraints and the full row rank assumption on P are satisfied.

3. Set $\hat{P}$ as $\hat{P} = D^{-1}\tilde{P}$ and $\hat{A}$ as $\hat{A} = \tilde{A}D$, where D is the diagonal normalizing constant matrix with each diagonal element being the row sum of $\tilde{P}$.

Remark 8: The consistency and asymptotic normality of $\hat{P}$ from CNLS2 can also be shown, though it is not provided in this article due to the limited space. Bootstrap standard errors can be obtained the same way as before.

Remark 9: The rank assumptions can be incorporated into the fitting procedure by forcing the condition numbers (the condition number of P can be defined to be the ratio of the smallest and the biggest eigenvalues of PP^t, and the condition number of A can be defined similarly based on A^tA) to be less than or equal to some threshold (say 30), i.e. in the optimization step, the feasible set of P (or A) is restricted to be the one satisfying the condition number constraints.

Remark 10: Note that the estimators from CNLS1 and CNLS2 are consistent even if the error variances are not constant (i.e. $\Sigma \neq \sigma^2 \mathbf{I}_p$ and Σ is unknown).

4. SIMULATION AND APPLICATION TO REAL DATA

In Section 4.1, we apply our method to the simulated data generated under model (5) with the assumptions B1–B2 and A1–A2, respectively. In Section 4.2, the method is illustrated with the Grand Canyon sulfate data.

4.1. Simulation

We first consider a simulated example to illustrate CNLS1 with B1–B3. The data are generated based on the model (5) with restrictions B1–B2, where $N = 150$, $n = 30$ (assuming the source contributions are repeated monthly), $p = 7$ and $q = 3$. Although the number of replications m_i need not be equal, for the sake of brevity, the same number of replications are used for the source contributions. Thus, $m_1 = m_2 = \cdots = m_{30} = 5$. The true source profiles (normalized to sum to 1) are given in Table 1. The

Table 1. True source composition profiles (P_0^t), estimated source composition profiles ($\hat{P}^t$) and bootstrap standard errors (BSE) for $\hat{P}^t$ from CNLS1 (using B1–B3)

Species	Source 1			Source 2			Source 3		
	True	Estimate	BSE	True	Estimate	BSE	True	Estimate	BSE
1	0.1347	0.1324	(0.0052)	0.1944	0.2043	(0.0091)	0.1529	0.1496	(0.0040)
2	0.1636	0.1675	(0.0040)	0.1409	0.1460	(0.0051)	0.1784	0.1725	(0.0036)
3	0.0210	0.0135	(0.0053)	0.1496	0.1354	(0.0138)	0.0113	0.0307	(0.0040)
4	0.1474	0.1550	(0.0046)	0.0757	0.0806	(0.0095)	0.1964	0.1817	(0.0050)
5	0.0053	0.0076	(0.0151)	0.1906	0.2036	(0.0186)	0.2180	0.2104	(0.0108)
6	0.3512	0.3389	(0.0153)	0.1958	0.1698	(0.0162)	0.0985	0.1254	(0.0119)
7	0.1768	0.1851	(0.0067)	0.0531	0.0603	(0.0136)	0.1445	0.1297	(0.0041)

Note: Bootstrap standard errors based on 50 bootstrap samples are provided in parentheses.

source contribution matrix A is generated from $U(0,3)$ with the conditions B1–B2 satisfied in Section 2. It is assumed that source 1 is missing on the 8th day, source 2 is missing on the 7th day and source 3 is missing on the 6th day, and when each source is missing, the average source contributions of the other sources stay the same. The errors associated with N observations are independently generated from the normal distribution so that the proportions of the error standard deviations to the model standard deviations are $\sim$23–27%. The resulting data matrix Y consists of nonnegative numbers. Note that the constrained minimization is done with A. Once we get the estimated source contributions, $\hat{A}$, the source profiles are estimated by the least squares formula, i.e. $\hat{P} = (\hat{A}^t M \hat{A})^{-1} \hat{A} M \bar{Y}$, $M = 5 \cdot \mathbf{I}_{30}$. Table 1 also contains the estimated source profiles and bootstrap standard errors for $\hat{P}^t$ based on 50 bootstrap samples. Although the nonnegativity constraints for the source profiles were not used, the estimates are all nonnegative. It is observed from the simulation that only when the true source profile matrix contains zeros are the corresponding estimates (of zeros) negative. In that case, it would be a natural choice to replace the negative estimates with 0 and renormalize each source profile. Figure 2 contains residual plots of $N(= 150)$ observations for each of seven species. Figure 3 shows the principal component plot of the data, the true source profiles and $\hat{P}$. It can be seen that $\hat{P}$ gives a very good approximation to the true source profile matrix.

The simulation is repeated 20 times. Since we are assuming that A and P are fixed parameters, the same A and P matrices are used for 20 simulations. Errors are regenerated from the normal distribution at each simulation. Throughout the simulation the proportions of the error standard deviations to the model standard deviations are $\sim$23–27%. The sample average and standard errors for $\hat{P}$ based on these 20 simulations are given in Table 2. Table 3 contains R^2 values between true source profiles and estimated profiles (R_p^2), and true source contributions and estimated contributions (R_A^2), respectively. It can be seen from these results that the estimated profiles (and contributions) are generally in very good agreement with the true source profiles (and contributions).

Secondly, we apply CNLS2 with A1–A3 to the data generated under model (5) with restrictions A1–A2, where $N = 150$, $n = 30$, $p = 7$, $q = 3$ and $m_1 = m_2 = \cdots = m_{30} = 5$. The true source profiles are given in Table 4. The source contribution matrix A is generated from $U(0,3)$ with conditions A1–A2 satisfied. It is assumed that source 1 is missing on the 8th and 15th days, source 2 is missing on the 7th and 14th days, and source 3 is missing on the 6th and 13th days. The errors associated with N observations are independently generated from the normal distribution so that the proportions of the error standard deviations to the model standard deviations are $\sim$22–27%. The resulting data matrix Y consists of nonnegative numbers. Note that the constrained minimization is

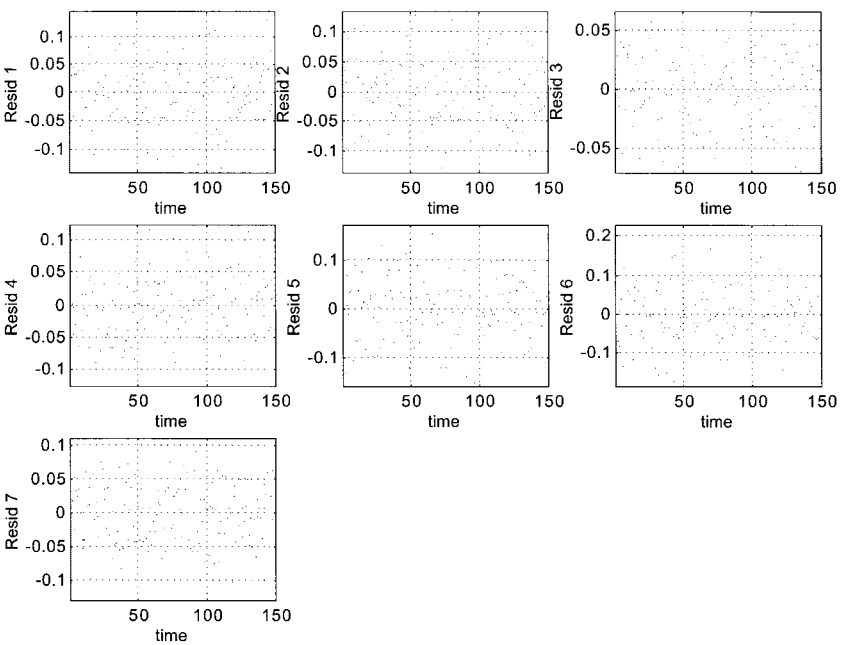

Figure 2. Residual plots of $N(= 150)$ observations for each species from applying CNLS1 with B1–B3

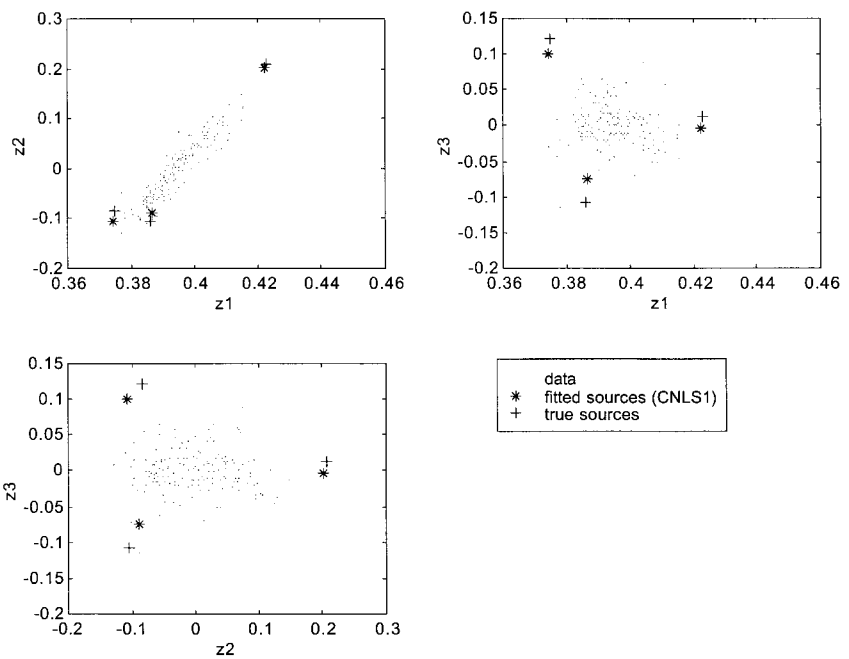

Figure 3. Principal component plots of simulated data (.), true sources (+) and fitted sources using CNLS1 (*) with B1–B3. Data and source profiles are normalized to sum to 1 before being plotted. The axes (z_i) are the eigenvectors of the cross-product matrix of the normalized data

Table 2. Sample average of $\hat{P}^t$ based on 20 independent samples (CNLS1 with B1–B3)

Species	Source 1	Source 2	Source 3
1	0.1327 (0.0042)	0.1903 (0.0043)	0.1578 (0.0038)
2	0.1654 (0.0039)	0.1443 (0.0051)	0.1740 (0.0036)
3	0.0139 (0.0028)	0.1338 (0.0040)	0.0298 (0.0035)
4	0.1526 (0.0033)	0.0906 (0.0042)	0.1804 (0.0048)
5	0.0103 (0.0089)	0.1952 (0.0134)	0.2141 (0.0073)
6	0.3458 (0.0112)	0.1835 (0.0120)	0.1093 (0.0071)
7	0.1792 (0.0036)	0.0623 (0.0049)	0.1345 (0.0036)

Note: Standard errors based on 20 independent samples are given in parentheses.

now done with P. Once we get the estimated source profiles $\hat{P}$, the source contributions are estimated by the ordinary least squares as given in point 1 of Remark 7. Table 4 also shows the estimated source profiles with bootstrap standard errors based on 50 bootstrap samples. Figures 4 and 5 contain the residual plots and the principal component plots, respectively. It can be seen that $\hat{P}$ gives a very good approximation to the true source profile matrix.

The simulation is repeated 20 times. As before, only errors are regenerated from the normal distribution at each simulation with $\sim$22–27% of the proportions of the error standard deviations to the model standard deviations. The matrices A and P are kept the same throughout the simulations. The results are reported in Tables 5 and 6. It can be seen from those tables that the estimated profiles (and contributions) are generally in good agreement with the true source profiles (and contributions).

Table 3. R^2 between P_0 and $\hat{P}(R_P^2)$ and R^2 between A_0 and $\hat{A}(R_A^2)$ over 20 simulations (CNLS1 with B1–B3)

Simulation no.	Source 1		Source 2		Source 3	
	R_P^2	R_A^2	R_P^2	R_A^2	R_P^2	R_A^2
1	0.9975	0.9862	0.9655	0.9555	0.9915	0.9443
2	0.9934	0.9848	0.9879	0.9657	0.9827	0.9311
3	0.9992	0.9913	0.9639	0.9495	0.9827	0.9325
4	0.9980	0.9927	0.9237	0.9276	0.9751	0.9389
5	0.9929	0.9901	0.9725	0.9573	0.9814	0.9340
6	0.9971	0.9888	0.9850	0.9494	0.9746	0.9513
7	0.9971	0.9890	0.9688	0.9498	0.9783	0.9456
8	0.9940	0.9891	0.9682	0.9378	0.9630	0.9426
9	0.9955	0.9893	0.9779	0.9518	0.9887	0.9445
10	0.9973	0.9802	0.9530	0.9460	0.9680	0.9437
11	0.9981	0.9856	0.9506	0.9613	0.9862	0.9315
12	0.9985	0.9835	0.9118	0.9683	0.9860	0.8850
13	0.9970	0.9894	0.9541	0.9530	0.9867	0.9410
14	0.9993	0.9786	0.9044	0.9638	0.9784	0.9175
15	0.9949	0.9885	0.9787	0.9519	0.9814	0.9424
16	0.9994	0.9689	0.8308	0.9494	0.9851	0.9081
17	0.9973	0.9879	0.9736	0.9352	0.9712	0.9339
18	0.9967	0.9882	0.9685	0.9423	0.9931	0.9504
19	0.9997	0.9888	0.9521	0.9515	0.9826	0.9430
20	0.9953	0.9880	0.9684	0.9576	0.9847	0.9381

Table 4. True source composition profiles (P_0^t), estimated source composition profiles ($\hat{P}^t$) and bootstrap standard errors (BSE) for $\hat{P}^t$ from CNLS2 (using A1–A3)

Species	Source 1			Source 2			Source 3		
	True	Estimate	BSE	True	Estimate	BSE	True	Estimate	BSE
1	0.0595	0.0657	(0.0119)	0.1086	0.1041	(0.0067)	0.2129	0.2194	(0.0039)
2	0.1586	0.1691	(0.0178)	0.2526	0.2504	(0.0079)	0.3573	0.3644	(0.0064)
3	0.0717	0.0889	(0.0169)	0.0043	0.0000	(0.0149)	0.2900	0.2988	(0.0062)
4	0.0732	0.0715	(0.0055)	0.1684	0.1592	(0.0073)	0.0223	0.0196	(0.0035)
5	0.2299	0.2063	(0.0155)	0.2464	0.2569	(0.0132)	0.0040	0.0000	(0.0026)
6	0.1201	0.1212	(0.0074)	0.2103	0.2085	(0.0075)	0.0827	0.0749	(0.0040)
7	0.2871	0.2773	(0.0192)	0.0094	0.0209	(0.0023)	0.0309	0.0229	(0.0039)

Note: Bootstrap standard errors based on 50 bootstrap samples are provided in parentheses.

4.2. *Application to real data: Air pollution spatial data*

The data set consists of 53 observations on 11 sites collected during the summer of 1992. Figure 1 contains the map of the 11 receptors used in this analysis. The variables are concentrations of a single species airborne particulate sulfate at 11 sites and the observations are 24 h average PM2.5 concentrations. The receptor model implies that the data can be represented by linear combinations of source profiles, which in this case represent the impact of several spatially distinct source areas. Though the regional pattern of sulfate concentrations would depend on the shifting wind patterns, the regional wind patterns are fairly constant over the same period as the observed air pollution

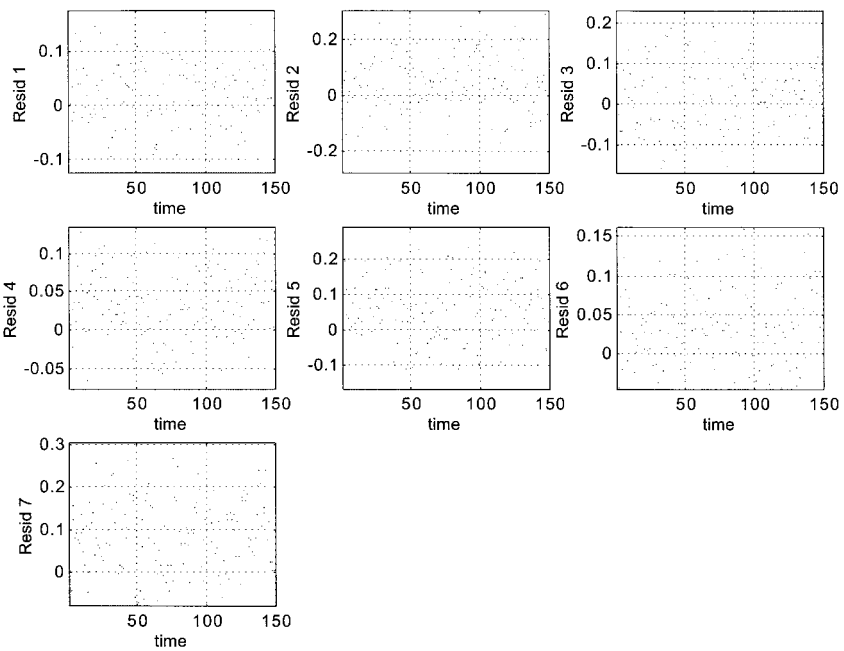

Figure 4. Residual plots of $N(= 150)$ observations for each species from applying CNLS2 with A1–A3

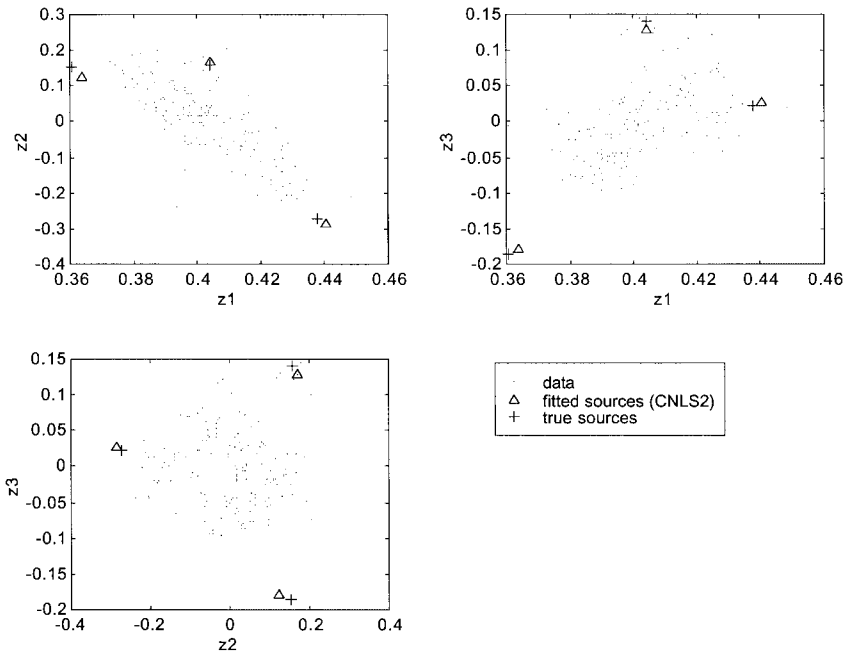

Figure 5. Principal component plots of the simulated data (.), the true sources (+), and the fitted sources using CNLS2 (△) with A1–A3. The data and the source profiles are normalized to sum to 1 before being plotted. The axes (z_i) are the eigenvectors of the cross-product matrix of the normalized data

concentrations. Regional wind patterns tend to be very consistent in any given season of the year. The main variability is between seasons. Of course, there is also large variability in regional wind patterns during the day that are caused by terrain. Examples of this are sea–land breezes and mountain valley flows. However, by using 24 h average data, these short time scale variations are averaged out.

As the referee pointed out, the spatial pattern of sulfate from each source is dependent on the wind patterns. Since there are only a few distinct source regions and only a few 1-day average regional wind patterns, then it is physically realistic to assume that there will only be a few underlying sulfate concentration patterns. Since the overall spatial wind flow patterns are approximately constant, so too the associated concentration patterns are approximately constant. Note that it is the spatial pattern or source profile for each that is relatively constant, not the concentrations.

Table 5. Sample average of $\hat{P}^t$ based on 20 independent samples (CNLS2 with A1–A3)

Species	Source 1	Source 2	Source 3
1	0.0241 (0.0153)	0.1227 (0.0083)	0.2112 (0.0066)
2	0.1096 (0.0206)	0.2685 (0.0083)	0.3570 (0.0071)
3	0.0259 (0.0216)	0.0416 (0.0167)	0.2856 (0.0100)
4	0.0827 (0.0097)	0.1480 (0.0088)	0.0244 (0.0057)
5	0.2806 (0.0204)	0.2137 (0.0147)	0.0057 (0.0082)
6	0.1253 (0.0095)	0.1945 (0.0092)	0.0849 (0.0055)
7	0.3517 (0.0306)	0.0109 (0.0089)	0.0312 (0.0126)

Note: Standard errors based on 20 independent samples are given in parentheses.

Table 6. R^2 between P_0 and $\hat{P}$ (R^2_P) and R^2 between A_0 and $\hat{A}$ (R^2_A) over 20 simulations (CNLS2 with A1–A3)

Simulation no.	Source 1		Source 2		Source 3	
	R^2_P	R^2_A	R^2_P	R^2_A	R^2_P	R^2_A
1	0.8791	0.9925	0.9100	0.9853	0.9973	0.9140
2	0.9607	0.9814	0.9886	0.9875	0.9992	0.9622
3	0.9348	0.9956	0.9699	0.9894	0.9981	0.9521
4	0.9409	0.9946	0.9670	0.9924	0.9998	0.9565
5	0.9456	0.9924	0.9028	0.9910	0.9995	0.9391
6	0.9178	0.9935	0.9428	0.9912	0.9995	0.9418
7	0.9067	0.9810	0.9283	0.9811	0.9929	0.9427
8	0.9261	0.9862	0.9678	0.9758	0.9942	0.9602
9	0.9501	0.9871	0.9637	0.9913	0.9961	0.9623
10	0.8806	0.9923	0.8558	0.9710	0.9971	0.8611
11	0.9937	0.9886	0.9882	0.9929	0.9992	0.9917
12	0.9091	0.9948	0.9110	0.9948	0.9994	0.9218
13	0.9061	0.9946	0.8538	0.9893	0.9996	0.8947
14	0.9362	0.9911	0.8850	0.9874	0.9993	0.9328
15	0.9978	0.9891	0.9982	0.9928	0.9993	0.9934
16	0.9456	0.9947	0.9445	0.9892	0.9984	0.9611
17	0.9416	0.9906	0.9595	0.9913	0.9990	0.9492
18	0.9884	0.9882	0.9901	0.9837	0.9969	0.9927
19	0.9537	0.9891	0.9666	0.9847	0.9994	0.9652
20	0.9363	0.9942	0.9503	0.9934	0.9995	0.9490

In this analysis, 53 observations themselves are regarded as $\bar{Y}$ in Section 3.2 since those measurements are daily averages of $PM_{2.5}$. Although measurement error variances are usually different for different chemical species, here we may safely assume that the error variances are the same, i.e. $\Sigma = \sigma^2 \mathbf{I}_p$ because only one chemical species, $PM_{2.5}$, was measured and meteorological conditions appeared to be approximately constant over the 11 sites. We apply CNLS1 with identifiability conditions B1–B3 to these data. For B1–B2, we assume that source 1 is missing on days 18 and 49, source 2 is missing on days 42 and 44, and source 3 is assumed to be missing on days 2, 16, 22 and 50, and also assume that when each source is missing, the average source contributions of the remaining sources stay the same. These edge points are determined by the plots (species by species plots and/or principal component plots) of the data and the SAFER fit results (Henry and Kim, 1990). The estimated source profiles and bootstrap standard errors based on 20 bootstrap samples appear in Table 7. The three estimated source profiles (spatial profiles) are closely related to the three largest known source regions: source profile 1 shows high (relative) concentration for the receptors, BARS and SAGO, which corresponds to transport from southern CA; source profile 2 shows high (relative) concentration for the receptors, GRCA and SYCA, which corresponds to particulate sulfate coming up from the smelters to the south; source profile 3 shows high (relative) concentration for the receptors, DOSP, MEAD, SPMO, ESSE and KING, which circle the large coal-fired power plant located near the center of the map (where CA, AZ and NV all come together). This interpretation of the spatial profiles is supported by a comprehensive analysis of a superset of this data using different techniques that were developed by Henry; see Henry (1997b).

We also apply CNLS2 using conditions A1–A3 to these data. For A1–A2, we only need to assume that source 1 is missing on days 18 and 49, source 2 is missing on days 42 and 44, and source 3 is assumed to be missing on days 16 and 22, which is weaker than B1–B2 in this case. Note that this is

Table 7. Estimated source composition profiles and bootstrap standard errors (BSE) for air pollution spatial data ($\hat{P}^t$), CNLS1 with B1–B3

Variables	Receptors	Source 1		Source 2		Source 3	
		Estimate	BSE	Estimate	BSE	Estimate	BSE
1	DOSP	0.0584	(0.0088)	0.0795	(0.0135)	0.1356	(0.0096)
2	GRCA	0.0125	(0.0129)	0.1442	(0.0237)	0.0443	(0.0060)
3	MEAD	0.0646	(0.0085)	0.0722	(0.0102)	0.1296	(0.0093)
4	MOSP	0.1443	(0.0134)	0.0163	(0.0208)	0.0965	(0.0091)
5	SAGO	0.2533	(0.0218)	0.0684	(0.0293)	0.0543	(0.0169)
6	SPMO	0.0609	(0.0076)	0.0716	(0.0109)	0.1355	(0.0106)
7	SQMO	0.0288	(0.0093)	0.1278	(0.0144)	0.0540	(0.0072)
8	SYCA	0.0097	(0.0086)	0.1574	(0.0295)	0.0207	(0.0076)
9	BARS	0.2583	(0.0264)	0.0486	(0.0288)	0.0842	(0.0138)
10	ESSE	0.0718	(0.0125)	0.1053	(0.0192)	0.1368	(0.0066)
11	KING	0.0375	(0.0114)	0.1086	(0.0073)	0.1084	(0.0043)

Note: Bootstrap standard errors based on 20 bootstrap samples are given in parentheses.

not always the case (see Remark 3). The estimated source profiles and bootstrap standard errors based on 20 bootstrap samples appear in Table 8. The results, in general, look close to those in Table 7. Figure 6 shows the principal component plot of the data and the fitted sources. From the plot again it can be seen that the estimated source profiles give a reasonable fit to the data.

5. CONCLUSIONS

This article has been concerned with consistent estimation of source profiles and contributions and uncertainty estimation. Although we presented the problem in the context of spatial data on a single species, the general methodology in the article is applicable to compositional data consisting of

Table 8. Estimated source composition profiles and bootstrap standard errors (BSE) for air pollution spatial data ($\hat{P}^t$), CNLS2 with A1–A3

Variables	Receptors	Source 1		Source 2		Source 3	
		Estimate	BSE	Estimate	BSE	Estimate	BSE
1	DOSP	0.0731	(0.0103)	0.0710	(0.0084)	0.1219	(0.0040)
2	GRCA	0.0298	(0.0058)	0.1287	(0.0117)	0.0568	(0.0049)
3	MEAD	0.0762	(0.0069)	0.0662	(0.0075)	0.1165	(0.0033)
4	MOSP	0.1245	(0.0084)	0.0318	(0.0174)	0.0880	(0.0058)
5	SAGO	0.2033	(0.0250)	0.1038	(0.0164)	0.0703	(0.0061)
6	SPMO	0.0744	(0.0084)	0.0643	(0.0077)	0.1208	(0.0043)
7	SQMO	0.0419	(0.0067)	0.1161	(0.0089)	0.0632	(0.0041)
8	SYCA	0.0249	(0.0069)	0.1419	(0.0138)	0.0401	(0.0051)
9	BARS	0.2097	(0.0213)	0.0824	(0.0139)	0.0910	(0.0065)
10	ESSE	0.0862	(0.0134)	0.0976	(0.0168)	0.1277	(0.0079)
11	KING	0.0560	(0.0100)	0.0963	(0.0058)	0.1036	(0.0032)

Note: Bootstrap standard errors based on 20 bootstrap samples are given in parentheses.

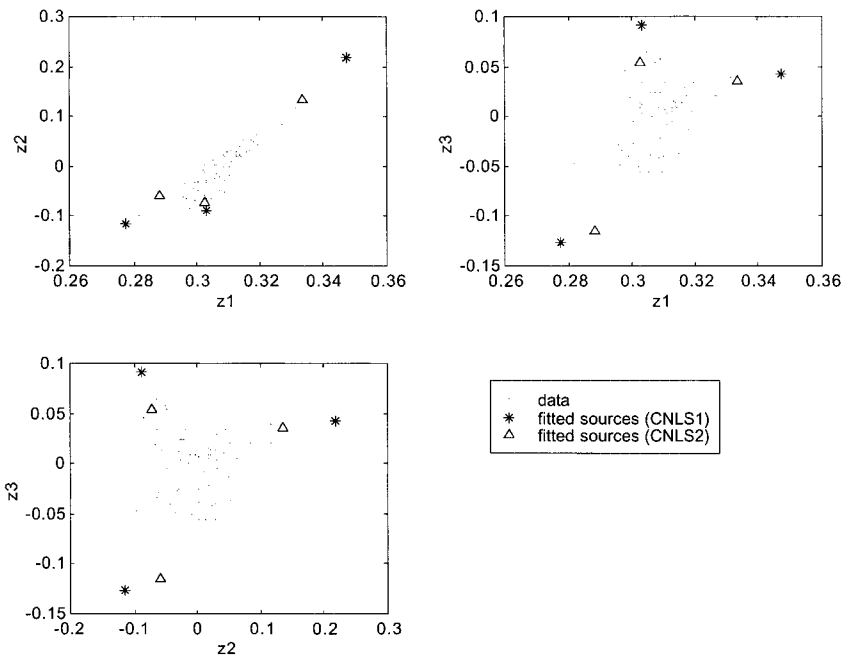

Figure 6. Principal component plots of the air pollution spatial data (.), and the fitted sources by CNLS1 (*) with B1–B3 and by CNLS2 (△) with A1–A3. The data and the source profiles are normalized to sum to 1 before being plotted. The axes (z_i) are the eigenvectors of the cross-product matrix of the normalized data

measurements on multiple species collected at a single receptor as well as a general factor analysis problem. To eliminate model nonidentifiability, new sets of identifiability conditions based on the source contribution matrix were proposed, which might be useful when the traditional identification conditions such as prespecification of 0s in the source profile matrix are not satisfied. The conditions based on the source contribution matrix are often realistic and usually require less prior information than the conditions based on the source profile matrix. Assuming the amount of errors is small, one might get an idea of where to assign zeros in the source contribution matrix by classifying the edge points of the data (if no prior information is available). The applicability of the proposed identifiability conditions is not restricted to the constant error variance structure nor the least squares method. They can be utilized in conjunction with MLE or Bayesian methods under a general error covariance structure if one is willing to make distributional assumptions.

Although it was not further discussed in the article, the conditions on the source contribution matrix can also be partially combined with the conditions on the source profile matrix to yield a complete set of model identifiability conditions when neither conditions serve as identifiability conditions by themselves (i.e. when neither A nor P has a sufficient number of zeros as considered solely).

The estimation procedures of the source profiles and the contributions are based on constrained nonlinear least squares. This approach has an advantage that no distributional assumption for the data needs to be made over the other parametric methods such as MLE. The resulting estimators are consistent as the number of replications goes to ∞. Standard errors for the estimators are provided by the bootstrap method.

ACKNOWLEDGEMENTS

Although the research described in this article has been funded in part by the United States Environmental Protection Agency (E. S. Park) through agreement CR825173-01-0 to the University of Washington, it has as not been subjected to the Agency's required peer and policy review and therefore does not necessarily reflect the views of the Agency and no official endorsement should be inferred. This research was partially supported by grant DMS-9523878 from the Chemistry and Statistics and Probability programs at the National Science Foundation (C. H. Spiegelman and E. S. Park) and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ESO9106).

APPENDIX A: PROOFS OF RESULTS

Assume $\text{rank}(A) = q$ and $\text{rank}(P) = q$ and A3 (B3) hold throughout. We need the following lemma to prove the results.

Lemma A1: Let r_{kj} denote the (k,j)th element of R, where $k = 1, \ldots, q$, $j = 1, \ldots, q$. Then

$$\sum_{j=1}^{q} r_{kj} = 1, \qquad k = 1, \ldots, q$$

Proof: We have $P = RP^*$ from $P^* = R^{-1}P$. Thus, the (k,j)th element of the matrix P can be expressed as

$$\sum_{i=1}^{q} r_{ki} p_{ij}^* = p_{kj}$$

Due to the constraint that row sum of P is 1,

$$\sum_{j=1}^{p} \sum_{i=1}^{q} r_{ki} p_{ij}^* = \sum_{j=1}^{p} p_{kj} = 1$$

and, by interchanging the summations,

$$\sum_{i=1}^{q} r_{ki} \sum_{j=1}^{p} p_{ij}^* = 1$$

It follows from the constraint $\sum_{j=1}^{p} p_{ij}^* = 1$ that

$$\sum_{i=1}^{q} r_{ki} = 1$$

A.1. Proof of Result 1

We also need the following lemma to prove Result 1.

Lemma A2: Under the assumptions A1–A2,

$$r_{kj} = 0, \qquad k = 1, \ldots, q, \quad j = 1, \ldots, q, \quad j \neq k$$

Proof: This can be proved with the same argument in Anderson (1984, pp. 576–577). Let the identifiability conditions A1–A2 hold. After reordering the rows of A, we can express A in the form

$$A = \begin{bmatrix} \mathbf{0} & A^{(1)} \\ \mathbf{a}_{(1)} & A_{(1)} \end{bmatrix}$$

where $\mathbf{0}$ is a zero vector of length $q-1$, $\mathbf{a}_{(1)}$ is a column vector of length $n-q$, $A^{(1)}$ is a $(q-1)$ by $(q-1)$ submatrix of rank $q-1$, and $A_{(1)}$ is the submatrix of $(n-q)$ by $(q-1)$. Let

$$R = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & R_{22} \end{bmatrix}$$

where R is nonsingular and r_{11}, r_{12}, r_{21} and R_{22} are of dimension $1 \times 1, 1 \times (q-1), (q-1) \times 1$ and $(q-1) \times (q-1)$, respectively. Also, let

$$A^* = AR = \begin{bmatrix} \mathbf{0} & A^{*(1)} \\ \mathbf{a}^*_{(1)} & A^*_{(1)} \end{bmatrix}$$

By a matrix multiplication,

$$AR = \begin{bmatrix} r_{21}A^{(1)} & A^{(1)}R_{22} \\ r_{11}\mathbf{a}_{(1)} + r_{21}A_{(1)} & r_{12}\mathbf{a}_{(1)} + A_{(1)}R_{22} \end{bmatrix}$$

Then r_{21} should be a zero vector since the rank of $A^{(1)}$ is $q-1$. Now do the same thing with the second column of A. This may be done after reordering the columns and rows of A (and also of R) so that the second column comes in first and the zeros are at the top. It follows that R should be a diagonal matrix.
Proof of Result 1: The conclusion immediately follows from Lemmas A1 and A2.

A.2. Proof of Result 2

We need the following lemma to prove Result 2.
Lemma A3: Under Assumptions B1–B2,

$$r_{kj} = 0, \qquad k = 1, \ldots, q, \quad j = 1, \ldots, q, \quad j \neq k$$

Proof: From $A^* = AR$, the (i,j)th element of the matrix A^* can be expressed as

$$\alpha_{i1}r_{1j} + \alpha_{i2}r_{2j} + \cdots + \alpha_{iq}r_{qj} = \alpha^*_{ij}, \qquad i = 1, \ldots, n, \quad j = 1, \ldots, q$$

and hence

$$\bar{\alpha}_1 r_{1j} + \bar{\alpha}_2 r_{2j} + \cdots + \bar{\alpha}_q r_{qj} = \bar{\alpha}^*_j \tag{A.1}$$

Say the kth $(k = 1, \ldots, q)$ source is missing on some days. Then

$$\bar{\alpha}^{(k)}_1 r_{1j} + \cdots + \bar{\alpha}^{(k)}_{k-1} r_{k-1,\,j} + \bar{\alpha}^{(k)}_{k+1} r_{k+1,\,j} + \cdots + \bar{\alpha}^{(k)}_q r_{qj} = \bar{\alpha}^{*(k)}_j \tag{A.2}$$

$j = 1, \ldots, q$ since $\bar{\alpha}^{(k)}_k = 0$. Here $\bar{\alpha}^{*(k)}_j$ is defined in a similar way as $\bar{\alpha}^{(k)}_j$. Subtracting (A.2) from (A.1),

$$\left(\bar{\alpha}_1 - \bar{\alpha}_1^{(k)}\right) r_{1j} + \cdots + \left(\bar{\alpha}_{k-1} - \bar{\alpha}_{k-1}^{(k)}\right) r_{k-1,\,j} + \bar{\alpha}_k r_{kj} + \left(\bar{\alpha}_{k+1} - \bar{\alpha}_{k+1}^{(k)}\right) r_{k+1,j} + \cdots + \left(\bar{\alpha}_q - \bar{\alpha}_q^{(k)}\right) r_{qj} = \bar{\alpha}_j^* - \bar{\alpha}_j^{*(k)} \tag{A.3}$$

By applying Assumption B2, we have

$$\bar{\alpha}_k r_{kj} = 0$$

for $j \neq k$. This implies that $r_{kj} = 0$ since $\bar{\alpha}_k \neq 0$ by the assumption that $\text{rank}(A) = q$.
Proof of Result 2: The conclusion immediately follows from Lemmas A1 and A3.

APPENDIX B: PROOF OF THEOREM

We need the following lemma to prove the theorem.
Lemma B1: Let $g(x, y)$ be a continuous real valued function defined on the Cartesian product $A \times B$, where A is a subset of p-dimensional Euclidean space and B is a compact subset of q-dimensional Euclidean space. Let x_0 be an interior point of A. Assume that the point y_0 is the unique point for which $\text{Min}_{y \epsilon B} g(x_0, y)$ is attained. Let $y_m(x)$ be a point in B such that

$$g(x, y_m(x)) = \text{Min}_{y \epsilon B} g(x, y)$$

Then $y_m(x)$ is a continuous function of x at $x = x_0$.
Proof: Appendix 4.B of Fuller (1987).
Proof of theorem: By the *WLLN*, as $m_i \to \infty, i = 1, \ldots, n$,

$$\bar{Y} \xrightarrow{p} A_0 P_0$$

Let $g(\bar{Y}, N^{-1}M; A) = \frac{1}{N}\text{tr}\left[\bar{Y}^t M \left\{\mathbf{I}_n - A(A^t M A)^{-1} A^t M\right\} \bar{Y}\right]$. Then, by the continuous mapping theorem,

$$g(\bar{Y}, N^{-1}M; A) \xrightarrow{p} g(A_0 P_0, C; A)$$

where

$$C = \begin{bmatrix} c_1 & & & 0 \\ & c_2 & & \\ & & \ddots & \\ 0 & & & c_n \end{bmatrix} \quad \text{and} \quad c_i = \lim_{m_i \to \infty} \frac{m_i}{N} > 0$$

Note that

$$\begin{aligned} \min_{\theta \epsilon \Theta} g(A_0 P_0, C; A) &= \min_{\theta \in \Theta} \left[\text{tr}\left\{P_0^t A_0^t C \left(\mathbf{I}_n - A(A^t C A)^{-1} A^t C\right) A_0 P_0\right\}\right] \\ &= \min_{\theta \in \Theta} \left[\text{tr}\left\{P_0^t A_0^t C^{\frac{1}{2}} \left(\mathbf{I}_n - C^{\frac{1}{2}} A (A^t C A)^{-1} A^t C^{\frac{1}{2}}\right) C^{\frac{1}{2}} A_0 P_0\right\}\right] \end{aligned}$$

is uniquely attained when $C^{\frac{1}{2}} A (A^t C A)^{-1} A^t C^{\frac{1}{2}} = C^{\frac{1}{2}} A_0 (A_0^t C A_0)^{-1} A_0^t C^{\frac{1}{2}}$ since the projection matrix is unique (see, for example, Rao, 1973, Section 1c.4). It follows that $A = A_0 R$ for any $q \times q$ nonsingular

matrix R. By the identifiability of the model parameters discussed in Section 2 (A1–A3 or B1–B3), this again implies that $A = A_0$. Thus, $\text{Min}_{\theta\epsilon\Theta}\, g(A_0P_0, C; A)$ is uniquely attained at $A = A_0$. By lemma B1, $\hat{A}$, which is the value of A such that $g(\bar{Y}, N^{-1}M; \hat{A}) = \text{Min}_A\, g(\bar{Y}, N^{-1}M; A)$, is a continuous function of $(\bar{Y}, N^{-1}M)$ and the result follows from the continuous mapping theorem.

REFERENCES

Anderson TW. 1984. *An Introduction to Multivariate Statistical Analysis*, 2nd edn. Wiley: New York.

Bandeen-Roche K, Ruppert D. 1991. Source apportionment with one source unknown. *Chemometrics and Intelligent Laboratory Systems* **10**: 169–184.

Bates DM, Watts DG. 1988. *Nonlinear Regression: Analysis and Its Applications.* Wiley: New York.

Davison AC, Hinkely DV. 1997. *Bootstrap Methods and Their Application.* Cambridge University Press: New York.

Efron B. 1982. *The Jacknife, the Bootstrap, and Other Resampling Plans.* CBMSNSF Monograph 39. SIAM: Philadelphia.

Fuller WA. 1987. *Measurement Error Models*. Wiley: New York.

Gleser LJ. 1983. Functional, structural and ultrastructural errors-in-variable models. *Proceedings of the Business and Economics Section.* American Statistical Association: Washington, DC; 57–66.

Gleser LJ. 1991. Measurement error models. *Chemometrics and Intelligent Laboratory Systems* **10**: 45–57.

Henry RC. 1991. Multivariate receptor models. In *Receptor Modeling for Air Quality Management*, Hopke P (ed.). Elsevier: Amsterdam; 117–147.

Henry RC. 1997a. History and fundamentals of multivariate air quality receptor models. *Chemometrics and Intelligent Laboratory Systems* **37**: 37–42.

Henry RC. 1997b. Receptor model applied to patterns in space (RMAPS) Part II—Apportionment of airborne particulate from Project MOHAVE 1997. *Journal of Air and Waste Management Association* **47**: 220–225.

Henry RC, Kim BM. 1990. Extension of self-modeling curve resolution to mixtures of more than three components. Part 1. Finding the basic feasible region. *Chemometrics and Intelligent Laboratory Systems* **8**: 205–216.

Henry RC, Lewis CW, Hopke PK. 1984. Review of receptor model fundamentals. *Atmospheric Environment* **18**: 1507–1515.

Henry RC, Park ES, Spiegelman CH. 1999. Comparing a new algorithm with the classic methods for estimating the number of factors. *Chemometrics and Intelligent Laboratory Systems* **48**: 91–97.

Hopke PK. 1985. *Receptor Modeling in Environmental Chemistry.* Wiley: New York.

Hopke PK. 1991. An introduction to receptor modeling. *Chemometrics and Intelligent Laboratory Systems* **10**: 21–43.

Kiefer J, Wolfowitz J. 1956. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics* **27**: 887–906.

Moran P. 1971. Estimating structural and functional relationships. *Journal of Multivariate Analysis* **1**: 232–255.

Park ES, Spiegelman CH, Henry RC. 1999. Bilinear estimation of pollution source profiles in receptor models. *Technical Report 019*. University of Washington, National Research Center for Statistics and the Environment.

Rao CR. 1973. *Linear Statistical Inference and Its Applications*, 2nd edn. Wiley: New York.

Spiegelman CH, Dattner S. 1993. Multivariate chemometrics, a case study: applying and developing receptor models for the 1990 El Paso winter PM_{10} receptor modeling scoping study. In *Multivariate Environmental Statistics*, Rao CR (ed.). Elsevier: Amsterdam; 509–524.

Yang H. 1994. Confirmatory factor analysis and its application to receptor modeling. Unpublished Ph.D. Dissertation. University of Pittsburgh, Department of Mathematics and Statistics.

